Ancestral State Reconstruction and Loanword Detection

Marisa Koellner
Seminar für Sprachwissenschaft
Universität Tübingen
marisa.koellner@uni-tuebingen.de

Johannes Dellert
Seminar für Sprachwissenschaft
Universität Tübingen
jdellert@sfs.uni-tuebingen.de

***Abstract*—Building on established applications of methods from bioinformatics to historical linguistics, we investigate the potential of different ancestral state reconstruction (ASR) methods for the task of loanword detection. Based on a very simple criterion for deriving loanword judgments from reconstructed ancestral states, we compare the performance of two state-of-the-art approaches to ASR against a very simple threshold-based, more linguistically motivated reconstruction method. We evaluate on the Indo-European cognacy judgments encoded in the IELex database. While overall performance is very low due to the properties of the dataset, there are marked differences in precision between the three methods, demonstrating that the development of specialized reconstruction methods for computational historical linguistics is worth pursuing.**

I. Introduction

Computational phylogenetics has developed a number of mathematical models and algorithms for calculating with evolutionary scenarios. Based on parallels between biological and linguistic evolution [1], [2], these new methods are increasingly being adapted to historical linguistics.

Ancestral state reconstruction (ASR), i.e. the reconstruction of states at the internal nodes of a tree based on attested states at the leaves, is an established component of approaches to phylogenetic inference [3]. The inference of cognate classes for unattested proto-languages in historical linguistics can be seen as equivalent to reconstructing the genomes of ancestral species in phylogenetics.

This makes reconstruction of ancestral states a valuable building block for **loanword detection** methods. In any such method, the performance of loanword detection will hinge on the quality of the hypothetical cognate classes at the internal nodes, which makes it relevant to compare different reconstruction algorithms in this application. Can the established methods from phylogenetics be applied directly, or is a more specialized reconstruction method needed to account for the peculiarities of the linguistic case?

In this paper, we propose a very simple specialized reconstruction method for cognate sets. To compare its performance to the two dominating paradigms of ASR in bioinformatics, we apply a simple loanword detection algorithm, and discuss precision and recall on the IELex database.

II. Data

The *Indo-European Lexical Cognacy Database* IELex [4] is currently the only lexical database which systematically covers an entire language family, and also contains expert judgements of both cognate classes and loanword status. For our evaluation, we use the IELex data for 207 concepts across 95 Indo-European languages.

In IELex, cognate classes are represented by multi-state characters at the leaves of the tree. The task of ancestral state reconstruction is then to postulate cognate classes at the internal nodes. Since we only want to compare reconstructions without considering the consequences of an imperfect phylogeny, we use an expert tree based on the classifications from *Ethnologue* [5] for living languages, and *Glottolog* [6] for extinct languages.

III. Reconstruction Algorithms

The traditional methods for ancestral state reconstruction in bioinformatics build on **maximum parsimony**, based on the intuition that the most likely reconstruction is the one which forces us to assume the least number of mutation events, i.e. lexical replacements in our case. Minimum parsimony can be seen as a formalization of Occam's razor, i.e. the principle of selecting the simplest hypothesis which explains all the data. To represent the state of the art in this paradigm, we use the variant of the Sankoff algorithm which is implemented in the PAUP* [7] software. In this implementation, the presence or absence of each cognate class at each internal node is inferred separately by minimizing the number of gain and loss events which need to be assumed.

The other important class of ASR method is based on **maximum likelihood** estimation. In this fully probabilistic approach, the states y at internal nodes are treated as unknown parameters whose values need to be estimated given the data x we observe. The maximum likelihood estimator maximizes $P(y|x)$, the probability of different parameter values given the observed data, by means of Bayes' rule. In the application to ASR, optimization is based on an explicit parametrized evolutionary model which fully describes how each state is likely to evolve along a given phylogenetic tree, and thereby assigns a probability $P(x|y,\theta)$ to the observed data for each

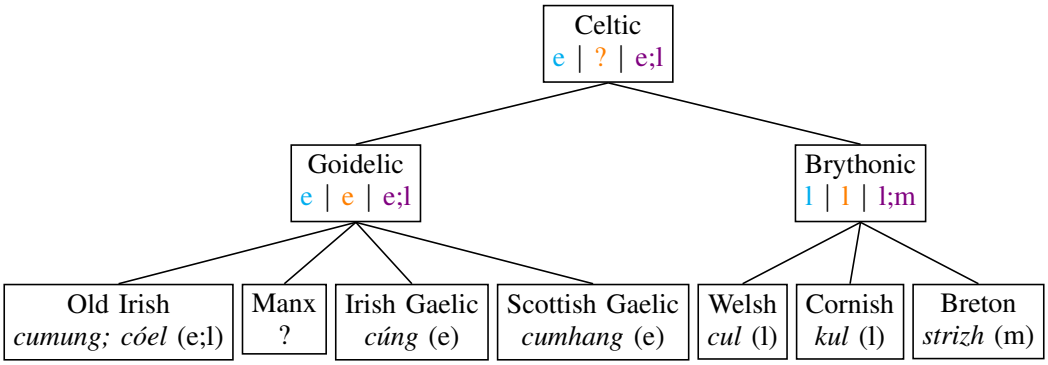

Fig. 1. The three reconstruction methods: Maximum Parsimony, Maximum Likelihood, Threshold-based

set of parameter values θ.

For our reconstruction, we use the marginal maximum-likelihood estimator implemented in the `phangorn` package [8], which is currently the most popular implementation of ML-based ASR. For our non-binary expert tree, branch lengths were estimated using PAUP* based on a model with constant rates of change without the molecular clock assumption, and empirically determined equilibrium distributions of the character states 0 (absence) and 1 (presence). The probability values for these states returned by `phangorn` at each internal node where turned into discrete reconstructions by reconstructing the cognate class in question whenever the probability of 1 was higher than 0.5.

As a third alternative, we test a hitherto unpublished **threshold-based reconstruction** method which is inspired by the way in which the presence of cognate classes in proto-languages is inferred in historical linguistics. This method only propagates evidence upwards in the tree, and is not built on any estimation of branch lengths, but exclusively on the number of child branches in which the cognate set is attested. To model these intuitions smoothly, we assign a confidence value $cn(v, c)$ to each cognate class c at each node v in the expert tree. For attested languages, the confidence value is defined to be 1 for each attested cognate class, and 0 for all other classes. The confidence values $cn(v, c)$ for non-leaf nodes are then computed recursively according to the following formula, where $Ch(v)$ is the set of children of v.

$$cn(v,c) := \max\left\{0, 1 - \frac{1 - \frac{\sum_{v_i \in Ch(v)} cn(v_i,c)}{|Ch(v)|}}{\sum_{v_i \in Ch(v)} cn(v_i,c) + 0.5}\right\}$$

This can be seen as a formalization of the following common types of argument in historical linguistics:

1) The presence of a cognate set in a proto-language of a family gets more likely with each branch of the family where the cognate set is attested.
2) If the ratio of child branches where the cognate set is attested is low, this somewhat detracts from our confidence that it was present in the proto-language.
3) If we have only two daughter branches and the cognate set is attested in only one of them, we still consider it more likely that it was already present at the mother, which we model by adding 0.5 to the denominator of the score.

We then reconstruct the cognate class c for a proto-language v whenever $cn(v, c) \geq 0.4$, where the threshold value was empirically determined by manual inspection of few examples.

To illustrate the differences between the three reconstruction algorithms, we inspect the resulting reconstructions of cognate classes for the concept "narrow" in the Celtic languages. Figure 1 shows the data in our version of the IELex database on the leaves, with the cognacy judgments encoded by the letters in brackets.

The Goidelic branch is dominated by the cognate class around Old Irish *cumung*, which is not present at all in the Brythonic branch. In that branch, the cognate class around Welsh *cul* dominates. This cognate class is also represented in Goidelic by Old Irish *cóel*. As an additional complication, Breton deviates by having *strizh*, a loan from a Romance language.

Both the maximum parsimony and the maximum likelihood method reconstruct only the one dominant cognate class for each branch, but they differ in the reconstruction for Proto-Celtic. The maximum-likelihood estimator does not assign a probability greater than 0.5 to the presence of any cognate class, which means that it does not have enough evidence for reconstructing any of the two classes for Proto-Celtic.

The Sankoff algorithm finds partial cognates for *cumung* in other branches of Indo-European (cf. German *eng* or Latin *angustus*), whereas cognates to *cul* are not attested anywhere else with the same meaning. This causes it to decide for reconstructing a cognate with *cumung* for Proto-Celtic.

The threshold-based method is very generous in assuming the presence of cognate sets at unattested nodes. For instance, it assumes that *strizh* was present in Proto-Brythonic, although

the word is only attested in Breton. The same holds for Proto-Goidelic, where the cognate class for *cóel* is reconstructed although the word is only present in Old Irish. In contrast to the two other methods, the evidence for *cumung* and *cul* is so strong on each branch that the confidence score exceeds the threshold for both classes in Proto-Celtic. This turns out to be the correct reconstruction in our case, which the other methods could not find because our version of the data did not contain the less common Welsh *cyfyng*.

IV. Loanword Detection

Our loanword detection procedure is based on a single very natural assumption: if a cognate class is reconstructed (or attested) for some node v, but a different class is reconstructed for its immediate ancestor, we mark all leaves under v as possibly having undergone borrowing. For each of these candidate leaves, we then determine whether the innovating cognate class is also present at another node in the tree. If this is the case, we conclude that we are indeed dealing with a borrowing.

Importantly, the use of this simple criterion implies that no attempt is made to detect borrowing from outside the family, and that the direction of borrowing remains underspecified. We are not attempting to achieve general-purpose loanword reconstruction, but can only detect borrowing events where cognate class was replaced by another, and which took place among languages which are modeled in the dataset.

V. Evaluation

We use the simple loanword detection procedure to determine whether our new reconstruction method is indeed a better choice than maximum parsimony and maximum likelihood reconstruction for projecting cognate sets back in time, and to see how much achievable performance depends on the quality of the reconstruction method.

In the absence of large etymological databases, the IELex database is the only reliable source we can use to evaluate the output of our loanword detection algorithm against expert judgements. The expert judgements in the database only consist of a simple binary annotation where 1 indicates loanword status, and 0 indicates either the absence of borrowing or incomplete data. This is problematic because there can be true loans which are not annotated as such in IELex. Moreover, the binary feature only allows us to compare the target languages, because the sources of borrowing are not modeled in any way. Since our simple method makes no effort to detect the directionality of borrowing, whereas the expert judgments in IELex are directional, a very low performance of the algorithm can be expected. Still, we can evaluate the reconstruction methods against each other.

As an example, in Table I we give the total number of languages with loans in IELex, and the number of such languages the different reconstruction methods implied for the concept "mountain". The maximum parsimony reconstruction implies the presence of loans for 38 leaves of the tree, i.e. there is a difference between attested cognate classes and the reconstruction of the immediate ancestor for almost half of the languages, a very problematic result. Maximum likelihood reconstruction reduces the number of detected loans to 10 languages, which means that it performs much better, but still severely overgenerates. In contrast, our threshold-based reconstruction only contains borrowing patterns for two leaves in the tree.

TABLE I
Languages with loans for the concept "mountain"

	True	Found	Correct
Threshold-Based	4	2	2
ML	4	10	2
Parsimony	4	38	3

TABLE II
Precision and Recall

	True	Found (% recall)	Correct (% precision)
Threshold-Based	863	1367 (23%)	196 (14%)
ML	863	2248 (32%)	273 (12%)
Parsimony	863	4532 (45%)	373 (8%)

In the IELex expert judgments, four languages are marked as affected by borrowing: English, Old English, Shughni (two of three cognate classes), and Gurbet Romani. Based on our reconstruction, English and Old English are correctly detected. The same holds for the maximum likelihood method. The maximum parsimony reconstruction additionally detects one of the loans into Shughni, but only at the price of a very large number of spurious results. Similarly, our reconstruction and the maximum parsimony method correctly identified the Cornish word for "to sew" as the only loan for that concept, whereas the maximum likelihood method led to an additional spurious loan.

In these and many other cases, even though our reconstruction detects fewer loanwords, it does so more consistently, leading to more accurate results. The overall figures are given in Table II. In our IELex subset, 863 loans are annotated. From the maximum parsimony reconstruction, the loanword detection method derives 4532 borrowing events, of which only 373 mirror the expert judgments in IELex. With the maximum likelihood method, 2248 borrowing events can be detected, whereas only 273 are identified correctly.

Our threshold-based reconstruction only leads to 1367 detected loan events, of which 196 are correct. While the two standard methods lead to many spurious loans and thereby an artificially high recall, they do so at additional cost to the already very low precision. The main conclusion is that the quality of the reconstructions has a strong influence on precision, where maximum likelihood improves precision by 50% compared to maximum parsimony, and our specialized reconstruction method improves it by another 17%.

Still, the results clearly show that even with a good reconstruction, the potential of purely cognate-class based loanword detection is very limited.

The most severe limitation is that we can only detect borrowing events within the language sample. As an example of external borrowing, we revisit the data for the concept "mountain". Our simple criterion detects the Old English and English forms as loans, because the corresponding cognate class is also present within the Romance languages. In the case of Shughni and Gurbet Romani, the innovating cognate classes only occurred once in the data, since the source languages are outside the language sample.
The other major reason for isolated occurrences of cognate classes within the data is semantic shift. For instance, the concept "head" has two realizations in German: *Kopf* and *Haupt*. *Haupt* is the inherited word, which nowadays only has a figurative meaning. *Kopf* is cognate to the English word *cup* [9], but has become the only word for "head" in common usage. This change of meaning cannot be detected by the algorithm, since *Kopf* and its cognate *cup* belong to two different concepts.

For these reasons, we cannot simply assume that each cognate class which is only present once in the data is likely to be a loanword. Operating only on cognacy judgments for a very small number of concepts across a single family, we have no way to distinguish loans from languages outside the sample from the effect of lexical replacements caused by semantic shift. In addition, an isolated occurrence of a cognate class can also be due to lack of coverage for the relevant group of languages.

Another reason for the low overall performance is that at the description level of cognate sets, it is impossible to detect borrowings within the same cognate class. A case in point is the English word *they*. According to the Oxford English Dictionary [10], the word is a borrowing from Early Scandinavian which replaced Old English *hīe*. The IELex database models this borrowing as an internal borrowing from Old Norse to English or Middle English. These kinds of internal borrowings are necessarily invisible to a cognate-based method. We see that at the desciption level of cognate sets, much crucial information is lost, making it impossible to detect some types of loanwords, which explains the very low overall performance.

VI. Conclusion

We have seen that ancestral state reconstruction is an important building block for automated loanword detection, and that the choice of ASR method has a strong influence on the precision of loanword detection. We have shown that a very simple linguistically motivated threshold-based method more consistently leads to good results than direct application of reconstruction methods from bioinformatics. For linguistic purposes, the development of specialized reconstruction methods is therefore beneficial, and our simple threshold-based approach is a first step in this direction.

However, the results also indicate that all current reconstruction techniques and our simple loanword criterion leave much to be desired. The problem of loanword detection is too complex to be solvable by considering and reconstructing the presence or absence of cognates. Some limitations, like lack of data on source languages, will always remain be a problem. Substantially larger databases covering multiple language families will be needed to develop methods which can begin to tease apart the phenomena of cross-family borrowings and semantic shift. Since the collection of data is a time-consuming task, experiments on real data should be complemented by experiments on large simulated datasets, which we are currently working on.

Also, much more advanced techniques which take phonetic values into account will be needed for inferring the source language and the directionality for each borrowing event. In this area, it seems worthwhile to explore reconstruction of phonological representations instead of cognate classes, then building loanword judgments on alignments between candidate loans and the reconstructed forms.

Acknowledgment

This work has been supported by the ERC Advanced Grant 324246 EVOLAEMP - Language Evolution: the Empirical Turn, which is gratefully acknowledged.

The authors especially thank Armin Buch and Gerhard Jäger for providing us with their version of the IELex data, and important hints about the implementation.

References

[1] Q. D. Atkinson and R. D. Gray, "Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513–526, 2005.

[2] R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson, "Mapping the origins and expansion of the Indo-European language family," *Science*, vol. 337, no. 6097, pp. 957–960, 2012.

[3] J. Felsenstein, *Inferring phylogenies*. Sinauer associates Sunderland, 2004.

[4] "Indo-European Lexical Cognacy Database," http://ielex.mpi.nl/, 2015.

[5] M. P. Lewis, "Ethnologue: Languages of the world – sixteenth edition," *Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com*, 2009.

[6] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank, "Glottolog 2.5," Leipzig: Max Planck Institute for Evolutionary Anthropology, 2015. [Online]. Available: http://glottolog.org

[7] D. Swafford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sunderland, Mass.: Sinauer Associates, 2002.

[8] K. Schliep, "phangorn: phylogenetic analysis in R," *Bioinformatics*, vol. 27, no. 4, pp. 592–593, 2011.

[9] *Etymologie der deutschen Sprache*. Dudenverlag, Mannheim, 2013, 5. Auflage.

[10] *OED online*. Oxford University Press, http://www.oed.com, Accessed November 30, 2015.